ASTRO-WISE
An Astronomical Wide-Field Imaging System for Europe

Edwin A. Valentijn[1] and Konrad Kuijken[1,2]

[1]Kapteyn Astronomical Institute, P.O.Box 800, 9700 AV Groningen, The Netherlands
[2] Sterrewacht Leiden, P.O. Box 9513, 2300 RA Leiden, The Netherlands
www.astro-wise.org

Abstract. With the new one square degree high resolution wide field imagers coming on-line in the near future, such as the 268 Mpix OmegaCAM at ESO's VST and the 360 Mpix MegaCAM at the CFHT, a new European-wide approach has been initiated to handle and disseminate the expected very large data volumes. For OmegaCAM both individual programs, including monitoring programs, and large sky survey programs are planned. Strict data taking procedures facilitate pipeline data reduction procedures both for the calibration and the science data. In turn, the strongly procedurized data handling allows European-wide federations of data-products. On-the-fly re-processing of archival data on the request of individual users with their own plugs-ins or newly derived calibrations sets are facilitated in an internationally distributed system. Compared to the classical more static wide-field image archives the newly designed system is characterized by a much more dynamical type of archiving.

1 Introduction

The data volume produced by the new generation of wide field imagers such as OmegaCAM at ESO's VLT Survey Telescope (VST, Paranal) and the Megacam at the CFHT (Hawaii) will be enormous. OmegaCAM, at a rate of 5 dithered exposures on a particular field in 30 minutes and with 300 nights per year of observing time, will produce over 30 Terabyte of raw data per year. This raw data volume contains roughly 10 Terabyte of calibration data and 20 Terabyte of raw science data. Data processing will then produce another 10 Terabyte of reduced science data and may create, with about 100,000 astronomical objects per OmegaCAM field of one square degree, enormous catalogues. Even the astronomical source lists of measured galaxy parameters can easily accumulate to 3-5 Terabyte per year!

Both the archiving of the data volumes and the processing of the image data go beyond the capabilities of personal work stations, which forces the user communities back to an old operational model of centralized nodes which host processors and storage media. The ASTRO-WISE project described below links the data centers set up in several European countries to support the current generation of wide-field survey instruments:

- The Netherlands, lead partner in the construction of the OmegaCAM instrument (NOVA/Groningen)

- France, European partner in the Megacam project (Terapix, Institute Astrophysique, Paris)
- Italy, lead partner in the construction of the VLT Survey Telescope (INAF/Naples) and partner in OmegaCAM (INAF/Padua)
- Germany, partner in OmegaCAM (Münich Observatory)
- ESO, who will operate VST and VISTA
- UK, lead partner in the construction of VISTA

The European Commission contributes 1.5 Meuro to this project through its FP5 RTD programme.

The aim of Astro-Wise is to set up a common system for processing, analysing and disseminating wide field imaging data. Internally, the centers will be connected via a shared database, while externally data products can be served to the AVO/VO networks.

There is an enormous variety of scientific research anticipated with the wide field imagers, ranging from 'fishing' special objects (moving, variable, or extreme in some colour index) out of millions, to statistical studies of large sample of objects, e.g. for cosmological shear research. The variety of scientific analysis of wide field data precludes the definition of single standard data products, and on top of this the enormous data volumes involved make it nearly impossible to re-process the whole data volume to achieve new releases with better code or improved calibrations data.

To provide the necessary flexibility to the astronomical end-user, Astro-Wise will provide an architecture that facilitates reprocessing of data as needed. It will have a peer-to-peer network between the national data centers, which each provide processing power and storage media, with full knowledge of what has been done at the other node. In this approach the processing of the data is viewed as essentially an administrative problem. The eventual goal is that the system administrates itself, so that when an end-user requests a particular data item, the system has full knowledge of how it has been derived, and how to rederive it if it is based on an 'obsolete' version of code or calibration data, with the result immediately being available at all nodes.

Furthermore, the machine handling of the large images and the big data volumes is non-trivial: particularly the pipeline data reduction, image comparisons and combinations, working with very large source lists, and visualization are all demanding tasks, even with modern hardware. The Astro-Wise consortium aims to share experience, build new tools and disseminate software for visualising and handling large image data.

Here, we describe the Astro-Wise survey system in terms of supporting the OmegaCAM project, which will be its first application, but it could as well host MegaCam data, or future infrared survey data.

1.1 Concepts

To face the data volume problems it is crucial to build an environment that provides, in a systematic and controlled manner, access to all raw and all cali-

bration data while keeping track of all processing and data products in a wide area network.

This environment should allow the astronomer to plan, modify and rerun the reduction and calibration pipelines to fit the particular needs that follow from the astronomical questions posed to the data. In addition, the environment provides systematic and controlled ways of running source extraction algorithms such that other astronomers could benefit from the obtained results. The archives should thus store the reduced data and source lists, or allow regeneration of these data dynamically. Because of the large data volumes and the limitations of local data centers, this archive must link different data centers, i.e. it must be a full-fledged federated database. Users at one data center can then profit from activities at other data centers, where new and possibly better calibrations have been built.

This dynamical archive continuously grows as more raw data enters the system and as more data reductions and calibrations take place. It can be used both for 'small' and for large science projects generating and checking calibration data and exchanging methods and scripts.

A key functionality is the link back from derived source data to the original raw pixel data, associated calibration files and *all* other data items that went into the result. This allows the user of the system to:

- verify the processing steps that have led to a certain product, and
- to qualify the product in terms of personal scientific exploration
- rederive the result with up-to-date calibration, thus providing the basic logistics for on-the-fly re-processing.

1.2 Example applications

In order to appreciate the above concepts in practical terms, we mention some practical applications:

(a) The VST is to be operated in service mode, and deep multi-colour exposures may be built up from data taken over many nights. Because all data will be accessible, and calibrated in a similar fashion, the optimal combination of data can only be done by selecting observations of a particular quality (quality information is a standard attribute to the archived data). Quality can sometimes only be assessed at the final stages of data reduction, so the linking information back to the raw data uis necessary to build a homogeneous final survey input dataset.

(b) Facilitating source list production from well documented final survey images allows the astronomer to select sources on a 1 to 1,000,000 basis as true interesting and not spurious sources, for the quality of each individual source extraction is an integral part of the source properties. With the capability of extracting, in a homogeneous way, sources automatically from all reduced frames, variability studies (such as proper motions of asteroids or nearby stars, or just

flux variations), can be done easily.

(c) The archive system is the best place to monitor the instrument as all calibration files are meticulously administered. The trend analysis of instrumental properties becomes essentially a push-button operation.

(d) A database environment is also the perfect place to plan observations because one can get a convenient overview of the quality of existing data and plan for filling the gaps in the spatial and quality domain. Based on the already available information, the addition of data in other passbands, the increase in exposure time or requirements for better seeing conditions can be identified and translated into an observing plan. For large surveys, the feedback between the data reduction/archive stage and the observations scheduling is an important tool when creating homogeneous raw datasets.

All this is done in the continuously growing archive.

1.3 Philosophy

The system that should provide above functionality will not be geared to a single data product, but should be a flexible tool. In fact many observations done with VST/OmegaCAM will be made for specific projects, not explicitly part of an all-sky survey *per se*. To capitalize on this diversity, flexibility is essential: even while OmegaCAM is performing individual observing programmes with specific science objectives, much of these data can as well be used for other research programmes with different objectives. Per year of operations the camera will survey the equivalent of 1000 square degrees in 5 passbands. So after a few years of operations the archives will contain a considerable fraction of the Southern sky, in effect delivering the data for large area sky surveys. When the quality and sensitivy of these obervations are recorded accurately, then these data can be well used for addressing for example statistical studies, in a similar way as radio luminosity functions can be well derived from radio surveys, in spite of the varying sensitivity over the field of view of the primary beam of radio telescopes.

Overall, the environment should optimise the interaction between users and their data, giving the user easy access to all aspects (attributes) and processing (pipelines) of the data. This, ever growing, dynamic archive will be geared to optical (IR) wide-field image data.

2 The ASTRO-WISE Survey System

To efficiently archive and handle the data volume, the OmegaCAM data acquisition, calibrations and pipeline reductions are strictly procedurized. These procedures are integrated in the design of the pipeline data reductions. Thus the design of calibration and scientific data reduction procedures has focussed on developing standard observing scenarios. It uses object-oriented methods to implement the associated data reduction procedures.

2.1 Procedurizing

The two major components of the data taking are the scientific and the calibration observations. Both need to be procedurized and the associated observations should be performed automatically. This can be achieved by carefully defining observing modes and observing strategies that cover all observational conditions both for building a homogeneous survey and for doing arbitrary scientific observations. The ESO operations model on Paranal naturally allows such a strategy to be defined,

The next sections give an overview of these modes and strategies.

Observing modes The CCDs in a detector mosaic do not fill the focal plane completely. The basic technique to overcome any gaps or artifacts in the CCD pixels is to take more exposures of the same field with slightly shifted field centers (dithering) and to co-add the images off-line in the pipeline process. This same technique can also be used to filter out cosmic ray hits on the detectors.

We distinguish the following observing modes:

Dither has offsets matching the maximum gap between CCDs, ~ 400 pixels. It will be operated with N pointings on the sky, where $N = 5$ is the standard. Although this mode erases all the gaps in the focal plane and maximizes the sky coverage, the context map (which relates each pixel in the combined image to the original exposures) will be very complex. An advantage is that in this mode the photometry among the individual CCDs can be coupled quite robustly.

Jitter has offsets matching the smallest gaps in CCDs ~ 5 pixels. It is the mode that optimises the homogeneity of the context map and will be used during observations for which the wide CCD gaps are not critical. In this mode all the data from a single sky pixel originate from a single chip.

Stare allows reobserving one fixed pointing position multiple times. It is the main workhorse for monitoring the instrument and allows detection of optical transients.

SSO is the mode for observing Solar System objects. It has non-siderial tracking. Pipeline processing support for this mode is limited to the standard image processing — specialized techniques are required to extract sources from such images.

Observing strategies An observing strategy employs one or a combination of the basic observing modes. It also defines a number of additional instructions for the scheduling of the observations. The observing strategy will be recorded in the FITS headers of the observations. Optionally, this header information can be used in data reduction pipelines, particularly those operated by the Consortium when addressing the combination (e.g. stacking) of images. It is not expected that the ESO pipeline will recognize strategies, as the standard ESO pipeline will not combine various runs.

We distinguish between the following strategies:

standard which consists of a single observation (observation block),
deep which does deep integrations, possibly taken at selected atmospheric conditions over several nights,
freq which frequently visits (monitors) the same field on timescales ranging from minutes to months and has overriding priority on the telescope schedule,
mosaic which maps areas of the sky larger than 1°.

2.2 Processing

The observing modes and strategies are fully integrated with the data reduction software. Their precise definition and limited number make it possible to design an associated *data model*, in the form of classes, that drives the pipeline design for data reduction and calibration. Figure 1 and 2 represent an overview of the datamodel which connects the datataking at Paranal, various quality control operations, the derivations of the calibration data and the "image-pipeline" which transforms the raw images into photometrically and astrometrically calibrated images.

Once the data operations, types and classes are defined the pipeline design is relatively straightforward. We discriminate between a *calibration pipeline* producing and qualifying calibration files, often involving a trend analysis, and an *image pipeline* that operates as a black box. By passively applying the calibration files (CalFiles) the *image pipeline* transforms the raw data into astrometrically and photometrically calibrated images. At ESO headquarters these pipelines will run under the Data Flow System pipeline infrastructure. At the national data centers these pipelines will run in an integrated environment where all data and data reduction steps are archived. Because algorithms for data reduction in the optical wide-field imaging arena are well established we can concentrate on other aspects of the data reduction scheme. We can view the pipeline as *an administrative problem*, where most attention should be paid to what ancillary information should be available when.

Calibrations The *calibration pipeline* is the collection of tools specifically designed to obtain all required calibration files (Calfiles). The requirements for these tools are specified as baseline requirements on OmegaCAM calibrations (Valentijn et al 2001). The calibration plan for the VST includes a comprehensive overriding photometric program. At the moment we have identified about 35 requirements, ranging from "check the focus" to "determine and monitor the atmospheric extinction". Each of these requirements are fulfilled by dedicated procedures both for the data acquisition at the telescope and for the calibration pipeline, which produces calibration files. Furthermore, these processes will also result into go-no-go flags. In fact, with the settling of the baseline requirements and the calibration plan all 'classes' in the data reduction have been defined.

The objective is to have a minimal interdependence between these procedures. Thus, the calibration pipeline can run the various derivations of calibration files at various time scales, independent of the derivation of other calibration

files. For example, the derivation of the master bias CalFile could be done at a frequency of twice a week, master flat fields once a week, photometric zero point once a night and the cross calibration of filters once a year, with a minimum of interdependence between these processes. The execution frequency of the different procedures of the calibration pipeline is tied to the frequency of the corresponding observations. As a baseline, the various frequencies for different calibration observations are highly standardized. The creation mechanism of the CalFiles includes a time stamping module which, as a result of a trend and/or quality analysis, assigns a time range for which the CalFile is valid. The image pipeline recognizes timestamps.

Science observations The *image pipeline* transforms the raw science data into calibrated images and passively applies the calibration files (CalFiles) made by the *calibration pipeline* (in fact the image pipeline is used in calibration procedures where required). Thus, the image pipeline produces the calibrated science images, and together with the CalFiles, which were used to derive these images, sets the end product ready for the astronomer for detailed scientific analysis. Unlike the calibration pipeline, the image pipeline does not produce any CalFiles.

The descriptor data of the reduced science images are stored in the database. These descriptors contain a copy of all the FITS header items, but they also contain links to all the data items (i.e. objects) which were used to derive the particular result. The CCD pixel data are not stored in the database; instead, a reference to a frame is added to the descriptor.

The *image pipeline* has many steps. Although it is designed to function as an automated 'streamer', the intermediate results are stored in so-called SeqFiles, again containing FITS-headers, statistics, intermediate results and links to data items. The descriptor will be used to store data of persistent value, and references to the descriptors can be used to track input and output of the various pipeline operations. For example: SeqFile 636 (co-added image) will have a reference to a list of SeqFiles (SeqFile 633 –Astrometrically calibrated image), which were used as input. The descriptors of these SeqFiles can be used to determine, for instance, the distribution of seeings or zero points in the input data, even though the image data for these input images may no longer exist.

Share the load The huge amount of data that needs to be processed in a limited amount of time necessitates the use of high powered CPUs and large bandwidths. Due to the physical nature of the OmegaCAM camera a natural parallelism is introduced where frames from the 32 CCD's can be processed quite independently through major portions of the data reduction pipelines. The level of parallelization is rather coarse-grained and the implementation of choice is a Linux Beowulf cluster. Having a very large bandwidth for communication between the processing units is essential to allow rapid dissolving of data across the cluster.

Data storage with significant amounts of fast and local disk space (10 - 100 Terabyte), is needed to minimize network traffic and at later stages allow distributed storage of processed data.

The data reduction will go in two stages. First the calibration is derived and CalFiles produced, then the image pipeline is run (at speeds of at least 1 Mpix/s) to produce calibrated science images.

The storage media need to have enough room for the images created throughout the lifetime of the project. This amounts to several 100's Terabyte. The archival storage of source parameters, depending on the use of the system and the total number of users, can easily go beyond the 10 Terabyte level.

In the federated environment the network plays an essential role. In an ideal world there is no need for replication of data, when information stored at a remote data center is needed it is delivered at the time of processing to the processing unit. This requires sustained network connections of 200 Mb/s or better. Such networks are becoming reality in the academic world these days. Even if the network speed is below this critical limit however, a 5 Mb/s network allows full replication of all OmegaCAM data on the 24 hours per day basis.

2.3 Federation

All the I/O of the pipeline processes goes to a federated database. The federated database is *the* archive in Astro-Wise where *all* information regarding the data and the processing of OmegaCAM will be stored. First of all, the raw data are accessible from the archive. The raw data itself will not reside inside the database, but the description of the data, including its location, will be available. This means that the data can only be manipulated through interaction with the database. In fact the methods (or pipelines) for processing of all kinds are also part of the federated database. When these methods are executed, they will interact with the database making sure it correctly describes the state of the OmegaCAM data repository.

Next to the raw data, calibration results, reduced images and source lists (possibly in the form of catalogs) will also be stored in the federated database, either as fully integrated objects or as descriptors.

Concepts of the federation A federation is a database environment that is spread over different physical locations but maintains a single database in the true sense of the word. The consortium is currently building such a system using Oracle -9i with SQL and Python interface. The choice for Oracle was motivated because of its support for object oriented programming, its scalibility up to Terabyte regimes of partitioned tables (needed for source catalogues), its Advanced Replication component supporting our federation and its new component Streams which allows the system to connect to e.g. ESO's Sybase archive and Terapix's MySQL archives. Astro-Wise has been adopted into the company's reference programme, supporting "innovative and leading projects". Also, the

availability of these components today (Astro-Wise needs to be able to start operations and receive first data by early 2004) and the size of the company in a very competitive database world has played a role in our decision.

According to Oracle: "A federated database is a logical unification of distinct databases running on independent servers, sharing no resources (including disks), and connected by a LAN."

In this environment full history tracking of all input will be done. To this end we employ object oriented inheritance techniques, links to objects (references in Oracle speak) and the database support for persistent objects. So everything in terms of processing that went on producing a result is readily available. The same set of links and persistent objects should also provide the back-bone for the on-the-fly reprocessing. To tag data and attributes in this very dynamical archive, context areas are introduced in the object attributes. Objects in this terminology are the persistent forms of the Object Oriented programming objects (used in the Python scripting language) that are the software counterparts of all OmegaCAM entities (of which a number are displayed in Figure 1. Some of these contexts can be:

- Project, with possible values Calibration, Science, Survey, or Personal
- Owner, with possible values pipeline, developer, or user
- Strategy, with values Standard, Deep or Freq (monitoring),
- Mode, with values Stare, Jitter, Dither or SSO and
- Time, with time stamping.

These context areas can be used to partition off areas of the database for certain projects. It will allow individuals to maintain their own partition, but also for larger projects, like large Sky Surveys to maintain parts of the data with project wide defined levels of quality control. The context areas also facilitate public access and provide the mechanism to interface the database to public browsers such as envisioned by the Virtual Observatories.

The federated database makes the Object Oriented programming languages objects persistent. Therefore any creation of a persistent object in the pipeline automatically has a counterpart that will be stored in the database. Because all data processing (intermediate) products have been defined in the OmegaCAM data model, classes can be programmed in the Python scripting language. The Object Oriented inheritance is also available from the persistence implementation, usually in terms of object links. For each (persistent) class a number of methods are defined which directly interact with the federated database, thus insuring database integrity.

The object oriented scripting language Python is used throughout the project providing the glue between the different working environments, such as:

- access to the database, through SQL,
- various pipeline codings as built by the scientific programmers and
- scripts provided by the astronomer-user allowing them to run "own methods" provided it fits into the datamodel.

Actually, another way of putting this is "the system provides the possibility to operate user customized pipelines and still maintain persistency".

The actual implementation of the database connectivity from the Python scripting language allows for a 'file structure' implementation of the database environment as well, thus allowing the pipeline to operate on files in a directory structure, completely independent from a federated database. However, in this case many of the advantages of a global environment are lost.

The current planning is to have the system ready for data acquisition by the very end of 2003, to test and populate it in 2004 and to prepare it for further mass production for 2005 and beyond. The system will be deliverable to satellite nodes at other European locations.

References

1. Valentijn, E.A., Begeman, K.G.B., Boxhoorn, D., Deul, E.R., Rengelink, R., Kuijken, K.H., 2001: VST-SPE-OCM-23100-3050 OmegaCAM Data Flow System User Requirements, ESO Garching, www.astro.rug.nl/∼omegacam/documents
2. Valentijn, E.A., Deul, E.R., Kuijken, K.H., 2001; Preston, Observing and Datamining with OmegaCAM, in The New Era of Wide Field Astronomy, ASP Conference Series, Vol 232, p 392, eds. R.G. Clowes, A.J. Adamson and G.E. Bromage www.astro.rug.nl/∼omegacam/documents

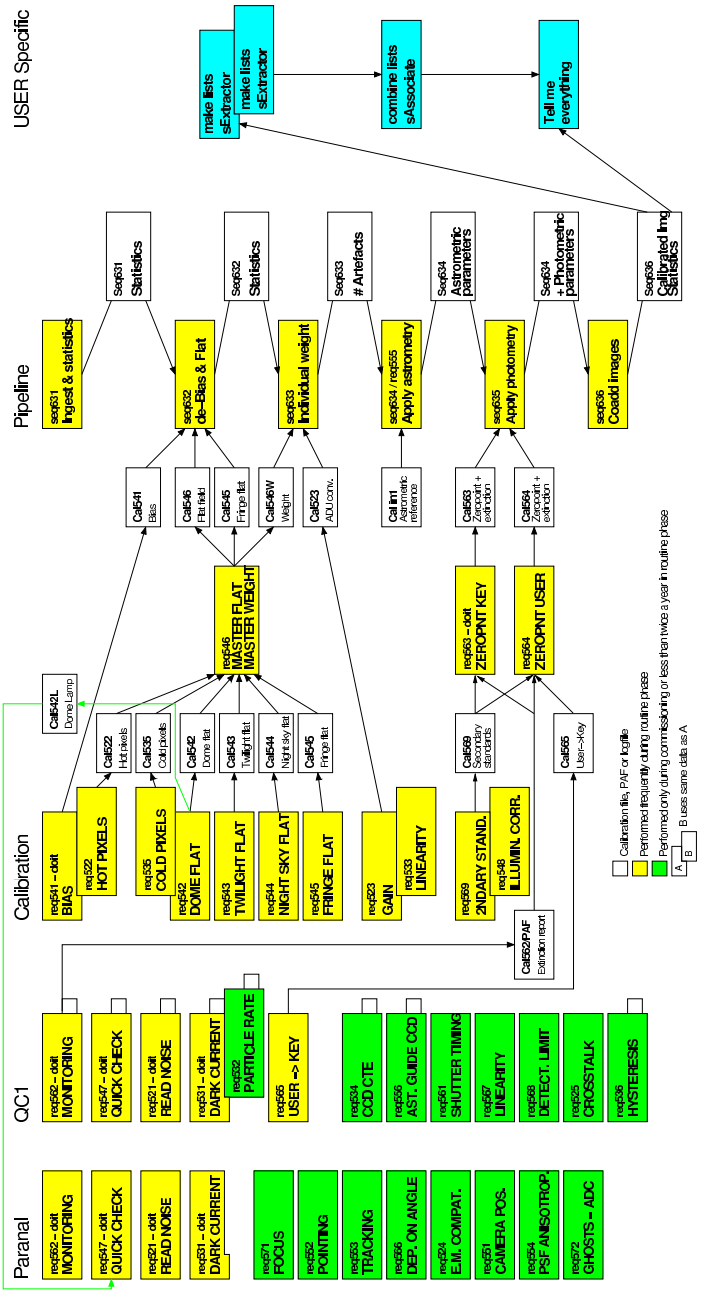

Fig. 1. The OmegaCAM datamodel with local quality control procedures highlighted in green (B/W dark) and persistent operations, ie. visible and re-runable by the end-user highlighted in yellow(B/W light)

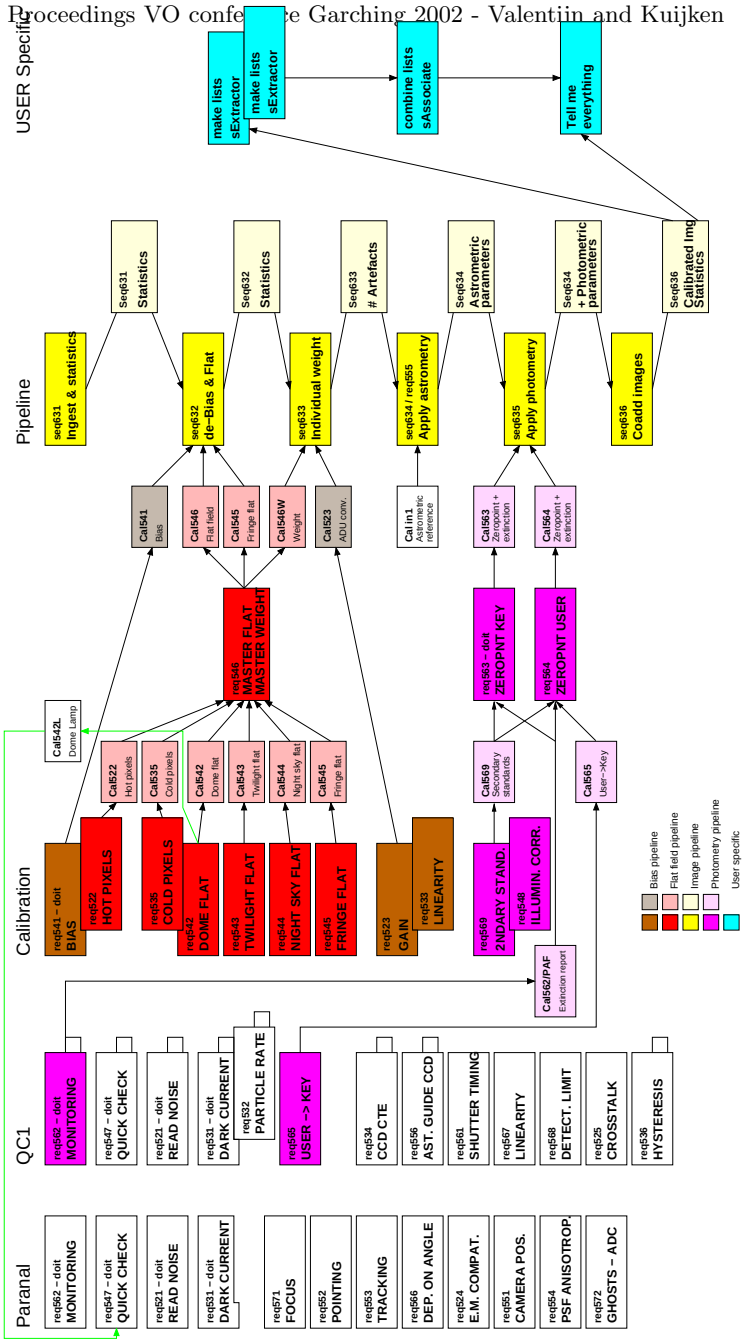

Fig. 2. The OmegaCAM datamodel with reduction pipeline procedures indicated by shades of gray.